ReVeS Participation - Tree Species Classification using Random Forests and Botanical Features

Guillaume Cerutti, Violaine Antoine, Laure Tougne, Julien Mille, Lionel Valet, Didier Coquin, Antoine Vacavant

ReVeS Participation - Tree Species Classification using Random Forests and Botanical Features*

Guillaume Cerutti[1,2], Violaine Antoine[4], Laure Tougne[1,2], Julien Mille[1,3], Lionel Valet[4], Didier Coquin[4], Antoine Vacavant[5]

[1] Université de Lyon, CNRS
[2] Université Lyon 2, LIRIS, UMR5205, F-69676, France
[3] Université Lyon 1, LIRIS, UMR5205, F-69622, France
[4] LISTIC, Domaine Universitaire, F-74944, Annecy le Vieux
[5] Clermont Université , Université d'Auvergne, ISIT, F-63001, Clermont-Ferrand

Abstract. This paper summarizes the participation of the ReVeS project to the ImageCLEF 2012 Plant Identification task. Aiming to develop a system for tree leaf identification on mobile devices, our method is designed to cope with the challenges of complex natural images and to enable a didactic interaction with the user. The approach relies on a two step model-driven segmentation and on the evaluation of high-level characteristics that make a semantic interpretation possible, as well as more generic shape features. All these descriptors are combined in a random forest classification algorithm, and their significance evaluated. Our team ranks 4th overall, 3rd on natural images, which constitutes a very satisfying performance with respect to the project's objectives.

1 Introduction

The ability to recognize a plant species and to understand its specificities has now become a task accessible mostly to specialists. Most flora books promise an arduous time to the willing neophyte, who does not possess the compulsory theoretical background. Mobile systems however offer the opportunity to introduce such knowledge in an interactive way, at the level of the user. Mobile guides for plant species identification have already seen light [1] with great success on white background images. The goal of the ReVeS project is to build a system to help users to recognize a tree in a natural environment, from the photograph of a leaf, in an educational and interactive way.

With this objective, we participated to the ImageCLEF Plant Identification task [9] for the second time, treating almost all of the 126 species in the database, making a strong distinction between the 24 species with compound leaves, and the 100 species with simple leaves we considered. The task consisted in associating, after a training phase, each one of the 3150 images in the Test database to an ordered list of species. Our work focused mainly of the application of methods dedicated to the case of photographs of one leaf in a natural environment.

* This work has been supported by the French National Agency for Research with the reference ANR-10-CORD-005 (REVES project).

2 Model Based Leaf Segmentation

Retrieving the leaf contour is the first and crucial step for the understanding of the image. It is a really challenging issue in unsupervised, complex, natural images [21,20] where it is necessary to incorporate as much knowledge as possible to ease the task [11]. Including prior knowledge on the expected shape of the object we look for is a good way to reduce the risk of mistakes. But in the context of a mobile application, it would be regrettable not to take advantage of a human user to guide an automatic process that would otherwise be very prone to erratic behaviour.

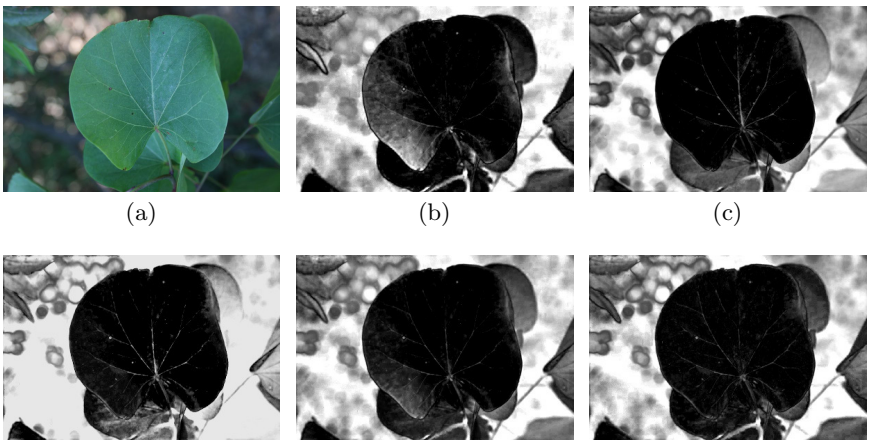
2.1 Leaf Color Model

In potentially imperfect images, and with the idea of approximating the shape of the leaf, color seems to be the most relevant information to rely on. A valid *a priori* color model for all leaves is impossible given the variety induced by season, species and lighting, but to extract a color model from every image, we first need to have a rough idea of the leaf's location.

What is quite easily solvable on white background image where a simple thresholding the grey-level image is enough to locate the leaf, becomes much more complicated with natural images. This is where we require the assistance of the user to draw a region inside the leaf, region that has to contain at least three components in the case of a compound leaf. We also rotated and cropped some photographs so that they clearly contain only one leaf of interest, with its apex pointing approximately to the top of the image, which corresponds to our frame of work for a mobile application. This is the only human intervention in the recognition process, and it is performed for photograph images only.

Fig. 1. Initial coloring on photograph images to locate leaf pixels

Based on this first, initial region, we try to estimate a model of the color of the leaf, and to compute the distance of each pixel to this model. This is done by using an evidence-based combination of the dissimilarity to a global color model computed by linear regression on the initial region (Figure 2(b)) and the dissimilarity to a local adaptive model built by adapting an expected mean color while exploring the image (Figure 2(c)). The use of belief function theory [6,18] is here more efficient than simpler combinations (mean, minimum) to take the specificities of two relevant yet distinct sources of information 2.



(a) (b) (c)

(d) (e) (f)

Fig. 2. Combination of dissimilarity maps of a leaf image (a) obtained from linear regression (b) and local mean (c) color models : using Dempster-Shafer theory (d) compared to the mean distance (e) and the minimum (f)

2.2 Active Models For Leaf Segmentation

We rely on explicit shape models to drive the segmentation of the leaves in images. It is however very complicated if not impossible to cover the diversity of leaf types, from needles to bipinnate leaves. This is why we had to introduce two different models, one for simple or palmately lobed leaves and one for compound leaves (pinnate, digitate and bipinnate indifferently) We thus considered only Angiosperm species, leaving *Ginkgo biloba* and *Juniperus oxycedrus* aside.

Parametric Active Polygon In the case of simple or palmately lobed leaves we still use a polygonal leaf model [5] to produce both a rough segmentation of the leaf and an estimation of its global shape. The number of lobes is estimated during the evolution, which is based only on the previously computed color dissimilarity map.

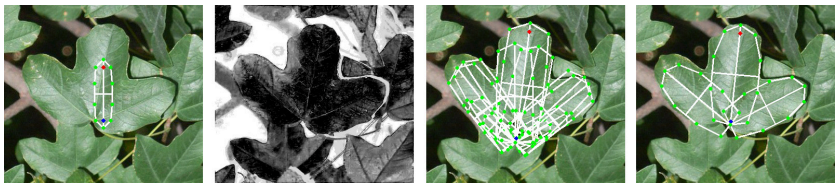
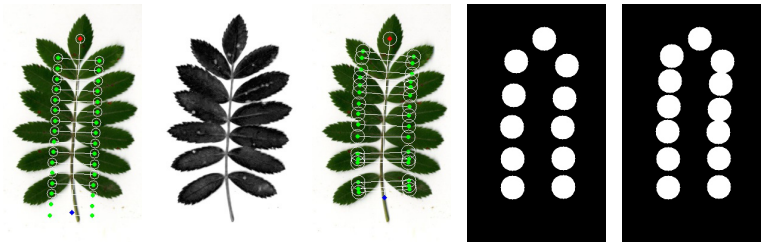


Fig. 3. Evolution of the parametric active polygon using the distance map, and redundant lobe suppression

A Model For Compound Leaves A leaf is supposed to be compound when the initial region (after either colouring or thresholding) has at least three connected components. To evaluate the number of leaflets n_F and their organization, we designed a deformable template for compound leaves that will evolve based on the dissimilarity map.



(a) (b) (c) (d) (e)

Fig. 4. Compound leaf model initialization (a) and evolution (c) based on the dissimilarity map (b); minimal and maximal estimations of the actual number and arrangement of leaflets (d, e)

This model represents leaflets by a fixed, excessive number of pairs of circles, and relies again on two points B and T and on a set of parameters to be optimized by minimizing an energy based on the total dissimilarity:

- k_F, the curvature of the axis
- r_F, the radius of the circles
- d_F, the distance of the circles to the axis
- $p_F(i)$, the relative position of each pair of circles

Given the possibility of overlap between leaflets, and subsequently model circles, we chose not to optimize the number of leaflets during the evolution, but to try to estimate it *a posteriori* with a a global view. Using the radius and position parameters, we first locate groups of connected circles, and based on their sizes and the gaps between them, we estimate how many leaflets actually compose each of them. This is a hard problem, and to reduce the risk of error, we make to estimation the retrieve a minimal and maximal number of leaflets and their location in the image. This process is illustrated in Figure 4. The numbers of leaflets and the parameters of the model are used as descriptors for the compound leaf structure.

To additionally represent the shape of the leaflets, we evaluate a simple polygonal model on one of the located leaflets. To minimize the risk of the model overflowing in the neighbouring leaflets and losing any accuracy, we chose the leaflet which stands out the most in the maximal leaflet estimation, considering the distance to the previous and next pair of leaflets.

2.3 Constrained Active Contour

We use the polygonal approximation both as an initialization and a shape constraint to retrieve the actual contour of the leaf, using an active contour algorithm for exact segmentation [5]. In the case of plain background images, the shape constraint is suppressed so that it doesn't prevent the contour to fit the actual leaf margin. Figure 5 shows the interest of including a shape constraint on complicated images.

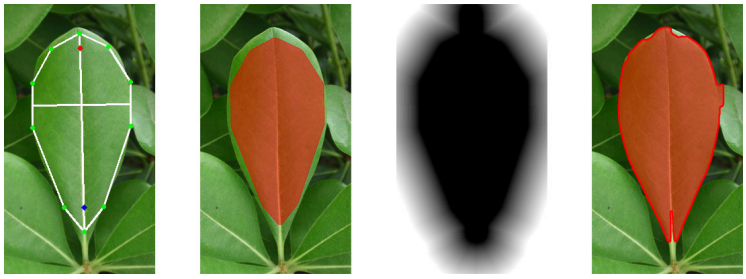


Fig. 5. Guided active contour results: polygonal model (left) initial contour (center left) guiding shape function (center right) and final contour (right)

3 Describing Leaf Shapes

To represent the discriminating properties encountered on the leaf, we chose to seek for the information investigated by botanists to identify species. The descriptors we use are then high-level morphological features designed to capture these specificities.

3.1 Global Shape Model

The represent the global shape of the leaf, we use directly the parameters obtained after the evolution of the global models. For simple and palmately lobed leaves, they consist of:

- model width w
- model center position p
- model apex angle α_A
- model base angle α_B
- number of lobes n_L
- length of each pair of lobes $l_L(i)$
- angle of each pair of lobes $\alpha_L(i)$

For compound leaves, we keep parameters from the compound leaf model, and those from the polygonal model extracted on one leaflet:

- minimal number of leaflets n_{Fmin}
- maximal number of leaflets n_{Fmax}
- position of the first leaflet h_F
- average gap between leaflets g_F
- distance to the axis d_F
- leaflet size s_F
- leaflet model width w_F
- leaflet model center position p_F
- leaf model apex angle α_{AF}
- leaf model base angle α_{BF}

3.2 Contour Interpretation

In order to capture more local shapes generally considered in the process of leaf identification, we rely on the axes of the polygonal model to partition the final contour into base, apex, lobe tip and right or left margin areas and know where to look at to estimate the determining features characterizing the leaf margin, basal shape and apical shape.

Curvature Scale Space Transform A rich representation of the contour used for shape matching is the Curvature Scale Space [13] that has already been used in the context of leaf image retrieval [12,4]. It consists simply in piling up curvature measures estimated on a normalized contour, producing a visually intuitive, scale invariant description of a contour. Figure 6 illustrates the interest of this transform for leaf margin analysis.

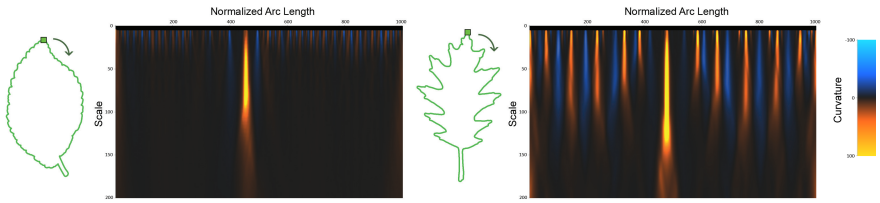
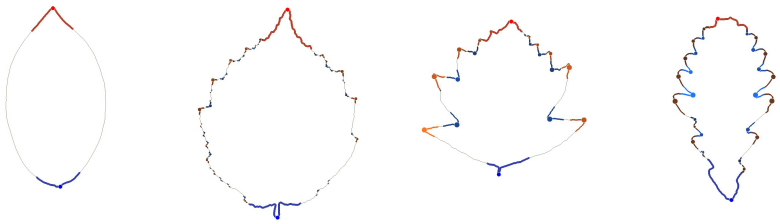


Fig. 6. Example of curvature-scale space transforms of two different contours (green square marks the first point)

CSS Image Description As such, the CSS image presents already a lot of visual information that may well represent the properties of the margin. Considering it as a describable visual object, we decided to extract texture information on this very image. We computed Haralick descriptors on a grey level version of the image to get a generic source of information on the margin, that does not take the risk of making too many assumptions.

Detecting and Characterizing Teeth On the other hand, we also wanted to locate and describe explicitly the teeth and pits on the leaf's margin. Considering that such structures clearly stand out in the CSS representation as maxima and minima of curvature, we followed an approach close to the detection of dominant points [19,14] to retrieve at each scale the salient features on the contour. For each one of them, the last scale at which a point is detected informs us on its size, while its sharpness can be estimated as the mean curvature of the point at all the scales it is detected.

Such points are searched only in the contour areas corresponding to the margin, so that the largest elements (apex, lobe tips, petiole) do not absorb the interesting smaller teeth. This detection process produces an interpretation of the contour where the base and the apex are precisely located, and with a sequence of convex and concave parts, characterized by their scale S and curvature K, as depicts Figure 7



(a) (b) (c) (d)

Fig. 7. Various leaf contours with detected base, tip, teeth and pits ; tip area in red, base area in dark blue ; convexities in orange, concavities in blue, brightness representing curvature intensity, extent representing scale.

To produce descriptors suitable for classification, we chose to compute the average values and standard deviations for both concave (−) and convex (+) structures, giving a set of 8 margin descriptors:

- mean scale of teeth $\bar{S}_+$
- standard deviation σ_{S+}
- mean scale of pits $\bar{S}_-$
- standard deviation σ_{S-}
- mean curvature of teeth $\bar{K}_+$
- standard deviation σ_{K+}
- mean curvature of pits $\bar{K}_-$
- standard deviation σ_{K-}

Additionally, we measured what percentage w_+ of the total margin length was part of a convex element, what percentage w_- was part of a concave one, and what percentage w_0 was part of none. These descriptors sum up some of the specificities botanists would look at to characterize a leaf margin, yet in a very condensed and efficient way.

3.3 Basal and Apical Shape Estimation

To account for the shape of the leaf int the basal and apical areas, we transposed the approach used with the global shape, by designing a simple, parametric, flexible model to adjust to the contour. It is attached to the contour point detected as the base or apex and composed of two parametric Bzier curves that try to minimize the distance of their points to the contour. Figure 8 shows examples of the evolution of these models.

The parameters used to build the necessary control points and optimized during the evolution, are also the descriptors we use to represent those shapes:

- α, the global angle of the model
- o, the orientation angle relatively to the leaf axis
- $(\alpha_{t1})_{l,r}$, for each side, the tangent angle at the tip point
- $(\alpha_{t2})_{l,r}$, for each side, the tangent angle at the end point
- $(\delta_{t1})_{l,r}$, for each side, the distance of the 1st control point from the tip
- $(\delta_{t2})_{l,r}$, for each side, the distance of the 2nd control point from the end

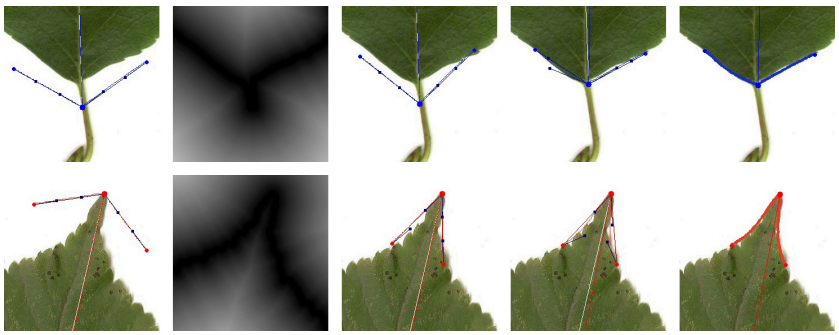


Fig. 8. Evolution of the base and apex models, based on the distance transform of the leaf contour

3.4 Statistical Shape Descriptors

In addition to these dedicated descriptors, we also consider more generic shape features that could be apply basically to any kind of object. The moments are a popular choice for their invariance properties and have already been used in the context of leaf images [21,12]. We computed the central moments on the segmented leaves and use them in combination with the other descriptors as a complementary shape representation, less directly pertinent but possibly more stable.

4 Feature Selection and Classification

Segmentation methods previously described enable us to build an ensemble of descriptors which defines the properties of the leaves. First, knowing this ensemble, we chose an appropriate classifier and we selected features that are strongly relevant for the classification, removing useless and noisy descriptors. We studied then the behavior of the selected classifier when its parameters vary. Plant identification was finally carried out using the optimized parameters of the classifier.

4.1 Particularity of the database

The extracted descriptors correspond to numerical properties of the images. It forms a vectorial data set that can be study in order to select a suitable classifier.

Let us first denote that the database may include noise, specially in the case of photographs, since the background images may contain colors quite similar to the leaves. Furthermore, the number of images per class is heterogeneous: some classes contain more than 200 objects whereas some are represented by less than 5 objects. This property can affect the efficiency of some classifiers. Finally, the difference between leaves for the same species are most of the time high. For example the shape or the color of a leaf can be different following its age or

depending on the season the picture has been taken. As a result, there exists a large variation of the similarity within species. Conversely, a low variation between classes may happen, as some species have similar leaves.

4.2 A random forest classification

Random forest, introduced by Breiman [3], is a recent technique which consists in ensemble of decision trees using for the final prediction a majority vote. To build a decision tree in a random forest, a bootstrap sample of the data is used and at each node a set of random variable is selected to split on. This random sampling strategy makes increase the error of each tree and reduces the correlation between trees. As a consequence, the ensemble achieves both low variance and low bias [3].

Random forest has shown high performances in many domains such as bioinformatics [10,15,7], ecology [16] or computer vision [17]. This method is robust to noise and generalize correctly the class models, even when the database has few examples per class. We decided then to use this classifier for the task of plant identification.

4.3 Methodology

For the next experiments, we used as decision trees the CART algorithm with the gini impurity measure [2]. We chose to built 200 trees and we set the parameter $mtry$, i.e. the number of randomly selected variables at each split, to $\sqrt{p}$ (where p is the number of attributes), as it is the default value commonly used in the literature. In order to measure the accuracy of the random forest, a 3-fold cross validation is employed.

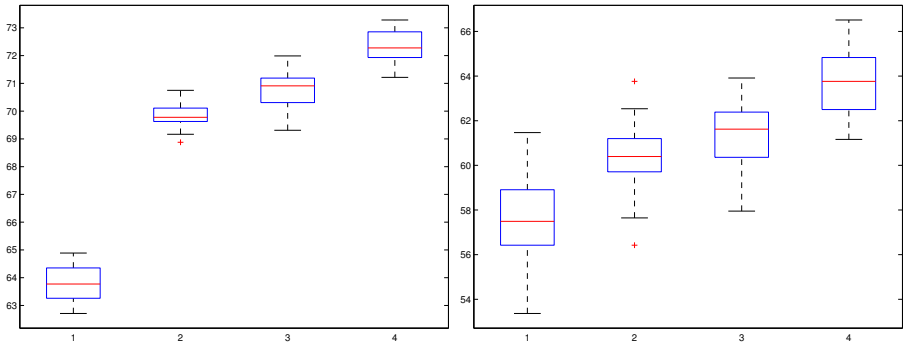
The dataset *Pl@ntLeaves II* includes three types of leaves: the first one corresponds to compound leaves, the second and third ones to simple leaves with and without lobes. For each type of leaves, the number of attributes differs, as well as the importance of the attributes in the classification process. Thus, we decided to built three random forests.

4.4 Feature selection

For simple leaves, we opted for the four following strategies:

- Strategy 1 consists in selecting the polygonal leaf model inducing the global shape model, the basal and apical estimations, as well as the attributes extracted from the CSS representation.
- Strategy 2 includes the polygonal leaf model, the basal estimation, the apical estimation and the contour characterization with the Haralick descriptors.
- Strategy 3 is a mix of the strategies 1 and 2.
- Strategy 4 includes the strategy 3 and the central moments.

Figures 9(a) and (b) present the results. We can remark that the Haralick descriptors applied on the contour of a leaf improve efficiently the classification. The strategy 4 corresponds to the best strategy for both types of simple leaves. We chose then this strategy for the rest of the experiments. Thus, 74 attributes are used for simple leaves with a unique lobe and 81 attributes for simple leaves with multiple lobes.



(a) (b)

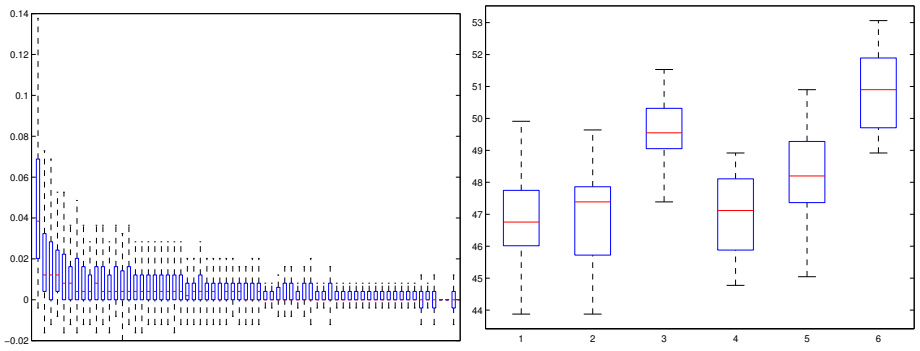
Fig. 9. Differences of accuracy for simple leaves with lobes (a) and without lobes (b) depending on the feature selection.

For compound leaves, the previous models (i.e. the polygonal model, the basal and apical model, etc.) characterize a small region of the image corresponding to one leaflet. As the resolution is lower than for simple leaves, the attributes are less efficient. We decided then to automatically erase useless or noisy attributes.

Such task can be performed using the variable importance measure described in [3]. This measure consists in observing the variation of the accuracy when one variable of the dataset is randomly permuted. Figure 10(a) shows the results when the attributes from all the models are taken. We can observe the existence of a quite large amount of useless variables, since they are represented by low measures. We decided then to remove all the variables that have an importance close to 0.

As a result, we considered 6 strategies which always includes the compound model, the polygonal model and the basal and apical shape estimations:

- Strategy 1 is composed of the Haralick descriptors computed from the contour detection.
- Strategy 2 corresponds to the strategy 1 with the add of the attributes extracted from the CSS representation.
- Strategy 3 includes the strategy 2 and the central moments.
- Finally, the three last strategies consist in the three first strategies suppressing the noisy variables.



(a) (b)

Fig. 10. Variable importance for compound leaves (a) and difference of accuracy depending on the feature selection (b).

Results are presented in figure 10(b). We can observe that deleting noisy variables leads most of the time towards better performances. We eventually chose the last strategy, ending up with 41 attributes.

4.5 Setting final parameters for the random forest

In order to achieve good performances, a random forest needs to fine-tune its parameters. One of the most important to adjust is $mtry$, the number of variables selected randomly at a split. Figure 11 presents the accuracy varying with $mtry$ for simple leaves with a unique lobe. When $mtry = 1$ the strength (i.e. the accuracy) of each tree is low and the correlation between trees is minimal. As a consequence, the ensemble performance is low. Conversely, if $mtry$ corresponds to the number of attributes, the decision trees of a random forest are built without randomness. Although the strength of each tree is high, the correlation between trees is quite important and induces low performances. Thus, we set $mtry = 20$ in order to have a good trade between strength and correlation. However, due to a lack of time when carrying out the experiments concerning this parameter, we chose for the contest to set $mtry = \sqrt{74} \approx 9$.

The same reasoning has been achieved for the two other types of leaves, resulting in a selection of 16 attributes for simple leaves with several lobes and 15 attributes for compound leaves.

A second important parameter to fine-tune is the number of trees to build for a forest. This number must be high enough to decrease the variance, leading to stable and accurate results. However, the time execution increases with the number of trees. We decided then to set this parameter to 5000 for the three forests.

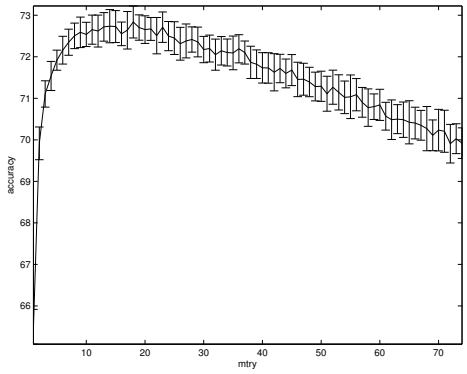


Fig. 11. Accuracy vs *mtry* for simple leaves with one lobe.

5 ImageCLEF Results

Three random forests are trained with the previous parameters and the whole dataset. As we only considered Angiosperm species, untreated leaves from the test set are assigned to *Ginkgo biloba* with a 100 percent confidence. For each leaf resting, a class probability is computed knowing the prediction of each tree of the forest. Results are presented table 1. Surprisingly, it shows better performances with pseudoscan images than scan images. This behavior has been observed by a quite important number of team. Thus, we suppose it is mostly due to the species presence in each type of images, as some class are more difficult to predict than others.

Table 1. Results using the average classification score of the contest.

Run	Scan	Pseudoscan	Photograph	Average
LIRIS_ReVeS_run_01	0.42	0.51	0.33	0.42

In total, 33 runs were submitted by 11 groups. We achieved 4^{th} place for pseudoscan images and the 5^{th} place for scan and natural images. As a team, we rank third team for photographs. This last result is a good performance, since we concentrated our effort on photographs and we used a poor human supervision.

6 Conclusions

The results we obtained are satisfactory enough to validate our general approach. The investigation of morphological features that model explicitly the criteria used by botanists to identify trees proves to be a convincing way to treat the problem of plant identification. The role of segmentation still seems prevailing,

and the performance on photographs, though being better than the average, underlines its importance.

A first implementation of our identification system on mobile devices has been engaged, with promising results, but not on as many species, leaving notably aside compound leaves. With the number of potential classes increasing, it will become a necessity to reduce the scope of the search, by making advantage of the GPS system that now exists in every smartphone. Fusing geographical information together with image based descriptors is the biggest challenge of our futur work. And knowing in advance which species are likely to be found in the geographical area where the user stands would be a decisive step towards a truly reliable identification.

The other prospect made possible by the use of high level descriptors will be the possibility of putting words on the extracted information. Explaining what the system understands and rendering it in a comprehensible form is a way to teach a non-specialist user what to look for, and could be used to implicate the user in the decision process by his intuitive corrections. All in all, such an explicative and interactive application would constitute a way not only to help recognize a plant, but to teach and transmit a rare knowledge.

References

1. P. Belhumeur, D. Chen, S. Feiner, D. Jacobs, W. Kress, H. Ling, I. Lopez, R. Ramamoorthi, S. Sheorey, S. White, and L. Zhang. Searching the world's herbaria: A system for visual identication of plant species. In *European Conference on Computer Vision*, 2008.
2. L. Breiman. *Classification and regression trees*. Chapman & Hall/CRC, 1984.
3. L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
4. Carlos Caballero and M. Carmen Aranda. Plant species identification using leaf image retrieval. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, CIVR 2010, pages 327–334, 2010.
5. G. Cerutti, L. Tougne, J. Mille, A. Vacavant, and D. Coquin. Guiding active contours for tree leaf segmentation and identification. In *CLEF (Notebook Papers/Labs/Workshop)*, 2011.
6. Arthur P. Dempster. A generalization of bayesian inference. *Journal of the Royal Statistical Society*, 30:205–247, 1968.
7. R. Díaz-Uriarte and S. De Andres. Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1):3, 2006.
8. H. Goëau, P. Bonnet, A. Joly, I. Yahiaoui, N. Boujemaa, D. Barthelemy, and J.-F. Molino. The imageclef 2012 plant identification task. 2012.
9. B. Goldstein, A. Hubbard, A. Cutler, and L. Barcellos. An application of Random Forests to a genome-wide association dataset: Methodological considerations & new findings. *BMC genetics*, 11(1):49, 2010.
10. A.-G. Manh, G. Rabatel, L. Assemat, and M.-J. Aldon. Weed leaf image segmentation by deformable templates. *Journal of agricultural engineering research*, 80(2):139–146, 2001.
11. F. Mokhtarian and S. Abbasi. Matching shapes with self-intersections: Application to leaf classification. *IEEE Transactions on Image Processing*, 13(5):653–661, 2004.

12. F. Mokhtarian and A.K. Mackworth. A theory of multiscale, curvature-based shape representation for planar curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14:789–805, 1992.
13. S. C. Pei and C. N. Lin. The detection of dominant points on digital curves by scale-space filtering. *Pattern Recognition*, 25(11):1307–1314, 1992.
14. J. Pooja and H. Jonathan. Automatic structure classification of small proteins using random forest. *BMC Bioinformatics*, 11(1):364, 2010.
15. A. Prasad, L. Iverson, and A. Liaw. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems*, 9(2):181–199, 2006.
16. F. Schroff, A. Criminisi, and A. Zisserman. Object class segmentation using random forests. In *Proceedings of the 19th British Machine Vision Conference (BMVC 08)*, Lake, UK, 2008.
17. Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.
18. C. H. Teh and R. T. Chin. On the detection of dominant points on digital curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11, 1989.
19. C.-H. Teng, Y.-T. Kuo, and Y.-S. Chen. Leaf segmentation, its 3d position estimation and leaf classification from a few images with very close viewpoints. In *Proceedings of the 6th International Conference on Image Analysis and Recognition*, ICIAR '09, pages 937–946, 2009.
20. X.F. Wang, D.S. Huang, J.X. Du, X. Huan, and L. Heutte. Classification of plant leaf images with complicated background. *Applied Mathematics and Computation*, 205(2):916–926, 2008.